Németh Márton

Webarchívum mint a tudományos kutatások tárgya

Egyre fontosabb bemutatnunk a hagyományos tartalomfejlesztési feladatokon túlnyúlva, hogy a webarchívum miként jelenhet meg a tudományos kutatások tárgyaként. Nagyon fontos felvillantani, hogy társadalmi szinten mi lehet az a hozzáadott érték, melyet a webarchívum kutatása kapcsán a felszínre kerülhet. A közgyűjtemények szempontjából nagyon fontos új perspektívákat kínál az új kutatási irányok megalapozása, külső partnerségi formák feltárása. A tudományos és társadalmi presztízs emelkedését kínálja intézményi szinten is egy-egy jól megalapozott projektben történő közreműködés. A felsőoktatási intézmények számára pedig fontos új szinergiákat tárulhatnak fel a digitális bölcsészetek, az információtudomány, illetve az informatika oktatásának-kutatásának kapcsán. Ez a tanulmány a magyar internetarchiválás kezdeteit, illetve annak tágabb szakmai környezetét bemutató Phd dolgozat munkálatai során született meg. Egyfajta bevezetést kínál nyújtani a webarchívumok különféle megjelenési formáiba a tudományos kutatások tárgyaiként. Mindezek előtt azonban arról a keretrendszerről ejtünk szót, mely összefogja a webarchívumokra fókuszáló kutatásokat

Tárgyszavak: *weblap; digitális dokumentum; digitális archívum; információtudomány; adatbányászat; adatelemzés; kutatás*

A WARCnet projekt

A webarchívumokra mint gyűjteményekre irányuló tudományos kutatásokat, illetve a webarchívumok gyűjteményeinek tudományos kutatási célú hasznosítását európai szinten a WARCnet projekt keretében fogják össze. Ennek finanszírozási kereteit dán forrásokból biztosítják 2022 végéig. Az Országos Széchényi Könyvtár (OSZK) részéről 2020 őszének végén, a második online konferenciát követően tudtunk csatlakozni a projekthez. Résztvevői európai nemzeti könyvtárak webarchiválást végző munkatársaiból, tudományos kutatóiból, egyetemi oktatást-kutatást végző személyekből (elsősorban kommunikációkutatók és történészek), illetve informatikus fejlesztők közül kerülnek ki. Négy munkacsoport keretében folyik a tevékenység. Az első munkacsoport a különböző intézményi webarchívumok gyűjteményeit érintő összehasonlító kutatásokra fókuszál, a második munkacsoport a több webarchívum által közösen épített nemzetközi gyűjteményekben rejlő kutatási, elemzési lehetőségeket tárja fel, a harmadik munkacsoport a webarchiváláshoz szükséges információtechnológiai fejlesztések területén mozog, a negyedik pedig a nyílt adatok menedzsmentjének webarchiválási vonatkozásait tartja szem előtt, beleértve a téma tágabb digitális közgyűjteményi vonatkozásait is (például webarchívumok integrációja integrált könyvtári rendszerekbe, full-text keresési funkciók fejlesztése).[1] A projekt fő koordinátora az aarhusi egyetem professzora *Niels Brügger*, az ő munkáját segítik a munkacsoportvezetők, akik egyben az irányító bizottság tagjai is. Sok egyéb mellett értékes információkat kaptunk például a Bajor Állami Könyvtár webarchívumát érintő gyűjteményfejlesztési irányelvekről, melyek jó mintául szolgálhatnak az OSZK saját hungarika alapú gyarapítási elvek megfogalmazásához is. De szóba kerültek még a jogi szabályozás európai példái, a kutatási célú gyűjteményi hozzáférés európai szabályozási környezeteinek összehasonlításával. A koronavírus járvány múltával lehetőség lesz pályázni rövid 3-5 napos kutatási célú szakmai tanulmányutakra, a projektben résztvevő partnerintézményekbe.

A projekt során a Belga Nemzeti Könyvtár munkatársa *Friedel Geeraert* készített e sorok szerzőjével interjút az OSZK tematikus COVID-webarchívum gyűjteménye kialakításának tapasztalatairól[2]. Ezt érdemes lehet majd összevetni a későbbiekben a többi elkészült interjúval, melyek például a Dán Nemzeti Könyvtár[3], illetve a British Library[4] munka-

társaival is közzétételre kerültek ezek már e témakörben.

A tanulmány felépítése

Először egy szinte napjainkban született új tudományágat a webtörténetírást mutatjuk be. A második témakörben a webarchívumra mint a digitális bölcsészeti kutatások tárgyára térünk ki, a webarchívumban tárolt nagymennyiségű adatkészletek tudományos kutatási célú felhasználásáról, illetve az archivált adatok vizuális megjelenítéséről lesz szó adatbányászati és adatelemzési megközelítésben.

1. Webtörténetírás

A webtörténetírás egy nagyon fiatal, igazából a 2010-es években kibontakozó tudományág. Önálló tudományos folyóirattal is rendelkezik már Internet Histories címmel, első számának bemutatkozó tanulmánya átfogó bemutatással szolgál e tudományos területről Niels Brügger és nemzetközi kutatócsoportjának segítségével.[5] A kutatások tárgya nagyon széleskörű témaköröket foglal magában. A webarchiválás és történeti kutatások viszonyával itthon is önálló tanulmány foglalkozik Kokas Károly és Drótos László révén a Digitális Bölcsészet című folyóirat debütáló számában.[6] Önálló rövid tanulmányban nemzetközi kitekintést is tettem a webtörténetírásról, a webarchiválás egyéb kutatási célú hasznosítási lehetőségeinek felvillantásával együtt.[7] A web első 25 éves történeti kontextusának felvázolását Niels Brügger végezte el.[8] Ugyanő mutat rá arra is, hogy a digitális anyagok kutatási célú felhasználása, a digitális bölcsészetek előtérbe kerülése mekkora hajtóerőt jelent a teljes bölcsész- társadalomtudományi terület vonatkozásában is.[9] A web múltjának tanulmányozása kulcsfontosságú a jelen fejlődési tendenciáinak értelmezéséhez is. Ez nem csupán a szűken vett világháló múltjára érvényes, hiszen az internet fejlődésének, politikai, gazdasági, társadalomtörténeti aspektusa is kulcsfontosságú a jelen trendjeinek elemzése szempontjából.[10] Az alábbiakban a terjedelmi korlátok miatt csupán rövid áttekintéssel szolgálunk a legfontosabb vonatkozási pontok felvillantásával.

A történészeknek ugyanúgy fontos szerepet kellene játszaniuk a webarchiválás kutatási célú felhasználása mögött álló archiválási intézményrendszer kereteinek meghatározásakor, mint ahogyan az a hagyományos levéltárak szervezeti rendje és munkafolyamatai esetében a 19. században történt. Ezt a fontos nézőpontot is önálló tanulmány tárja fel *Susanne Belovari* révén.[11] Fontos feladata a történészeknek a webarchívumok biztonságának vizsgálata abból a nézőpontból, hogy a múlt archivált tényeinek manipulálását, aktuális politikai célokra történő újraírását is meg kellene akadályozni. Itt fonódik össze egymással az IT-biztonság és a történelmi hitelesség szempontrendszere.[12]

A számítógépes világhálónak mint technikai infrastruktúra történetének, valamint a web mint kommunikációs és publikációs platform történetének tanulmányozása hangsúlyosan előtérbe kerül. Emellett egy adott személy, esemény, témakör, intézmény webes lenyomatának nyomon követésére is lehetőség nyílik. Érdekes példát kínál erre az első amerikai weboldal történetének rekonstruálása[13] vagy a brit egyetemek weboldalainak története.[14]

A webes információ nagyon gyorsan avul, a digitális műveltség, illetve a személyes információk kezelésének fontos eleme lenne a személyes archiválás módszereinek széles körű oktatása és alkalmazása.[15] A Networkshop 2020 című digitális könyvtári és informatikai konferencián 2020 őszén önálló workshopot szerveztünk a témakör megtárgyalására.

A fentebb tárgyalt témaköröket egészíti ki az archivált szöveges, illetve vizuális webes tartalmak, vagy akár adott webszerverek naplófájljainak tanulmányozása a gépi tanulás (machine learning) eszköztárával, illetve a nagymennyiségű adatok elemzésének módszereivel.[16]

A kutatás szintjeként megjelenhet egy egyedi webes fájl, vagy weboldal, illetve egy adott webhely, doméntartomány is. Legtágabb értelemben pedig a webes univerzumnak, mint olyannak a történetét szintén lehet vizsgálni. A Memento-protokoll segítségével több webarchívum archivált anyagai is összevethetővé válnak egy adott weboldalról, illetve témakörről, illetve ennek szükségszerű korlátait is feltárták már.[17, 18] Egy másik nézőpont, amikor azt vizsgáljuk, hogy ki és milyen céllal, hatókörrel folytat webtörténeti kutatásokat. Amikor kismennyiségű forrásanyagot, akár csupán egy meghatározott honlap történetét tanulmányozzuk egy adott szoftverkörnyezettel, speciális kutatási céllal, azt nevezzük Niels Brügger terminológiájával élve mikroarchiválásnak, melynek felhasználási módjait gazdag szakirodalmi háttér villantja fel.[19] Persze dolgozhatunk webhelyek egy adott gyűjteményével például valamilyen speciális szakterület webes

lenyomatát tanulmányozva. Ilyenkor már a makro szintről beszélünk. Arra is voltak kísérletek, hogy egy teljes nemzeti webtartományt tanulmányozzunk. Erre a legkorábbi 2000-es évek elejei francia példától napjainkig rengeteg esettanulmányt találunk például Dániából, Hollandiából, Horvátországból és Szlovéniából.[20]:A nemzeti domén tanulmányozásának speciális esetei közé tartozik a volt Jugoszláviának kiosztott .yu domén, amely mára már teljes mértékben el is tűnt az élő webről, s már csak webtörténeti módszerekkel vizsgálható. Erre *Anat Ben-David* tett munkatársaival együtt kísérletet.[21]. Különféle részletes módszertani összefoglalók is napvilágot láttak már a webarchiválás vizsgálatának módszereiről és szintjeiről.[22]

Jelentős kihívásként jelenik meg a töredékes mementók megjelenése, a hibásan archivált objektumok, illetve a webarchívum visszanézése során felmerülő megjelenési problémák. A történeti hitelesség kérdését veti fel az, hogy az OpenWayback megjelenítő program különféle idősíkokban készült mentési elemeket csúsztat a visszanézés során egymásra, illetve az is előfordulhat, hogy néhány interaktív elem az élő webről szűrődik be a mentett anyag megjelenítésébe. Az archivált állomány autentikus volta tehát a visszanézés során is sérülhet. Egy adott webhely akár új helyre is költözhet, s az eredeti címén adottesetben már teljesen más tartalom kap helyet.[23] Fontos megemlítenünk ebben az összefüggésben, hogy miután maga a webarchiválás eredendő módon egy összefüggő egészből csak töredékeket tud kiragadni, ez óhatatlanul hatással van a webtörténészek tevékenységére is. Azzal kell dolgozniuk, ami a rendelkezésükre áll, s adott esetben megpróbálni pótolni a múlt hiányzó darabkáit. Ebben persze semmiféle nóvum sincs, hiszen a hagyományos történeti források használata is általában hasonló dilemmákat vet fel. Ami mégis különlegessé teszi a webtörténetírói munkát a hagyományos történészi tevékenységhez képest, hogy a vizsgálódásaink módszertani háttere megtervezésének van egy speciális összetevője. Egyaránt tisztában kell lennünk a webarchívumban tárolt források jellegzetességeivel, a webarchiválás munkafolyamatának főbb jellemzőivel, illetve annak a hardver- és szoftverkörnyezetnek a sajátosságaival, illetve korlátaival melyek révén vizsgálódásainkat folytatni tudjuk.

1.1 A webarchívumok történeti kutatási célú elemzésének támogatása

A webtörténészek munkájának támogatására egyre inkább előtérbe kerül speciális munkakörnyezetek kialakítása. Egyre súlyosabb kihívásként jelent meg, hogy a történettudósok közül sokan nem szándékoznak mélyreható informatikai ismeretekre szert tenni, miközben számukra is biztosítani kellene a webarchívumokban tárolt anyagok kutatási célú elemzését.[24] Az ausztrál, az új-zélandi és a brit nemzeti könyvtárak webarchívumai és az Internet Archive az internet archiválásért felelős nemzetközi konzorciumhoz (IIPC-hez) benyújtott nyertes pályázatában az ehhez szükséges munkakörnyezet kialakítását tűzte ki célul. Nem voltak előzmény nélküliek ezek a munkálatok, mivel az Archives Unleashed projekt keretében már a később tárgyalt projektben is részt vállaló Internet Archive elkezdett kifejleszteni online munkafüzeteket, illetve felhő alapú gyakorlókörnyezetet az Archive-IT által gondozott gyűjteményekre irányuló kutatások segítésére elsősorban könyvtárosok, levéltárosok illetve digitális bölcsészeti kutatásokkal foglalkozó szakemberek számára.[25] A következőkben ismertetett projekt ezen túlnyúlik, mert olyan keretrendszert kínál, mely bármely IIPC tagintézmény saját webarchívumához szabványosan illeszthető lesz.

A könyvtárak, levéltárak, múzeumok digitális gyűjteményeinek használatához a GLAM Workbench szolgáltatási környezet eddig is biztosított eszközöket, gyakorlati témákat és leírásokat. Ehhez kellett hozzáilleszteni a webarchiválás témakörét. A cél, tehát annak bemutatása lett, hogy a történeti kutatások során felmerülő témákat a webarchívumokból nyert adatok elemzésének segítségével miként lehet újszerű nézőpontokból feltárni.[26] Ehhez úgynevezett Jupyter digitális munkafüzeteket (notebook) hoztak létre. Az elméleti hátteret tartalmazó szöveges útmutatók mellett részletes leírások is találhatók gyakorlati példákkal a különféle elemzési módszerek használatáról.[27] Az igazán újszerű megoldást azonban egy webböngészőbe integrált gyakorlófelület jelenti, melynek segítségével hozzáférhetünk egy adott webarchívumhoz és a gyakorlatban is kipróbálhatjuk, hogy miként lehet adatbányászati, illetve adatelemzési tevékenységeket folytatni, konkrét példákon keresztül körbejárni különféle felhasználási lehetőségeket. Sőt arra is van lehetőség, hogy e gyakorlófelületeken különféle előregyártott alkalmazásokat futtassunk le, s mérjük fel azok kimenetének felhasználási lehetőségeit. A fejlesztők szándéka szerint a Memento protokoll, valamint a webarchívumok által használt korábban ismertetett eszközök (Heritrix, Brozzler, OpenWayback, PyWB) használatával bármelyik nemzeti könyvtári webarchívum összekapcsolható a tananyagokkal.[28] Így

akár arra is lesz lehetőség, hogy össze lehessen hasonlítani a különféle webarchívumok archiválási módszereit, a metaadatmodelleket, illetve az azokra épülő szolgáltatásokat. Lehetőség nyílik az időbeli dimenzió tanulmányozására, hogy miként fejlődött egy-egy webhely, vagy egy-egy témakör mikor jelent meg a weben s hogyan gyarapodtak az ahhoz kötődő honlapok számszerűleg, illetve tartalmilag egyaránt.[29] A fejlesztők szándékai szerint ez az oktatási környezet a zárt hozzáférésű archív gyűjteményekhez is hozzáilleszthető lesz a már említett szoftveres háttér biztosítása esetén. A prototípus őszre készül el. Ezt követően az IIPC keretében tanfolyamokat terveznek a különféle archívumokhoz történő illesztés, illetve a tananyagok használatának elsajátítására könyvtárosok számára. Fel lesznek mérve a továbbfejlesztési lehetőségek is.[30] Remélhetőleg Magyarországon is sikerül majd bevonnunk ezt az újszerű oktatómunkakörnyezetet a saját magunk oktatási portfóliójába. Így az akkreditált közgyűjteményi tanfolyamokon újszerűen be tudnánk mutatni a webarchívumunk felhasználási lehetőségeit, valamint a kutatói közösség felé új kapcsolatokat lehetne építeni az újszerű kutatási lehetőségek bemutatása révén.

2. A webarchívum mint a nagy mennyiségű adatok forrása, az adattudományi kutatások tárgya

A webarchívumok nagy szövegkorpuszok tárházaként adattudományi projektek középpontjában is állhatnak. Számos ilyen projekt felmerül már a szakirodalomban az utóbbi évekből.[31] Az értékes adatok gyors feldolgozása, illetve visszakeresésének biztosítása egyre inkább előkerül a webarchívumok használata során. Ilyen adatok lehetnek a naplófájlok adatai, speciális tranzakciós (pl. geolokációs adatok) vagy különféle, az adott archív gyűjteményben tárolt szöveghez kötődő adattípusok is.[32] A webarchívumok a disztributív adatfeldolgozáshoz is segítséget nyújthatnak Apache Hadoop segítségével egy megadott alkalmazáskészlettel, adott platformon. *Lnenicka* és munkatársai[33] egy teljes munkafolyamatot vázolnak fel egy webes tartalombányászati alkalmazás fejlesztésére, s egy big data alapú archívum létrehozására, mely modern alkalmazáskörnyezetet használ (Python, PHP, JavaScript, MySQL, és felhőszolgáltatások). Felvázolják az architektúrát, a módszereket, az adatstruktúrát a weboldalak adatainak bányászására, disztributív feldolgozására, és big data alapú elemzésére. Új típusú együttműködés jöhetne létre ennek alapján a közgyűjtemények, a webarchiválással foglalkozó szakemberek, illetve az adattudósok között. A szerzők arra is felhívják a figyelmet, hogy big data alapú alkalmazások az adattárolás, feldolgozás, elemzés kapcsán kiegészíthetik a hagyományosan használt programok tudását, de semmiféleképpen sem helyettesíthetik őket! A részlegesen strukturált, illetve teljesen strukturálatlan adatkészletek számos kutatási célból vizsgálhatók. Összpontosíthatunk a webes tartalmakra, a tartalomhasználati adatok kinyerésére, illetve a webes szerkezeti elemek feltárására is. A tartalomfeltárás egyes webhelyek, illetve webhelycsoportok által közölt információk visszakeresésében segíthet. A fő cél az, hogy strukturált adatokat nyerjünk ki ezekből a tartalmi erőforrásokból. Ezek az adatforrások aztán integrálhatók szemantikailag hasonló adatelemekkel, valamifajta tartalmi hierarchia vagy tartalomintegráció alkotható meg a segítségükkel.[34]

A különféle strukturált, illetve részben strukturált adatkészletek kvantitatív alapú történeti elemzések tárgyául is szolgálhatnak. Ebben az esetben az adattudós kinyeri az adatokat a webarchívum gyűjteményéből, s segítséget nyújt a webtörténésznek azok elemzésében. A Niels Brügger és *Ralph Schroeder* által szerkesztett munka, mely első ízben nyújt reprezentatív képet a webtörténetírás különféle alkalmazási példáiról, számos ilyen projektet sorol fel.[35] Ilyen például a brit országdomént bölcsész és társadalomtudományi szempontokból vizsgáló BUDDAH projekt[36], melyet 2014−15-ben bonyolítottak le. 65TB-nyi anyag került 1996 és 2013 között begyűjtésre; a projekt célja az volt, hogy különféle hasznosítási formákat találjanak a begyűjtött hatalmas adatkészlet kapcsán. Ez az anyag nem tükrözi teljes egészében a .uk domén tartalmát. A begyűjtött adatelemeket az aratás dátumával rögzített időbélyegekkel látták el az archiválási folyamat során. A fejlesztők és a kutatók közös munkájának eredményeként megszületett a SHINE névre hallgató keresőfelület, mely a begyűjtött anyagban történő teljes szövegű keresést tette lehetővé. A visszakeresést segítette az anyag különféle témakörökre bontása is, mely a szabadszavas keresés mellett szintén a visszakeresés alapjául szolgálhatott. A webarchívumban tárolt anyag koncepcionális elemekké szervezése, a kutatási stratégiák felállítása, illetve a visszakereső eszköz, illetve navigációs felületének tervezése közben új együttműködési területek tárultak fel a különféle tudományágak képviselői között.[37] Számos kihívás persze továbbra is fennállt a projekt lezárását követően. Hogyan kezeljék a nem teljeskörűen, illetve zavaros tartalommal archivált

adatelemeket, a webarchiválásra fókuszáló kutatási irányok hogyan illeszthetők be az egyes hagyományos tudományterületek keretei közé, a kutatás során felmerülő kérdéseket hogyan lehet közérthető módon bemutatni. A történeti típusú kutatásokra szolgáló keresőmotor prototípusát nyilvánosan is elérhetővé tették.[38]

Egy újabb érdekes alkalmazási területet villant fel az idén év végén záruló LinkGate nevű projekt, mely a webarchívumokban tárolt nagymennyiségű adatok vizualizációjával foglalkozik.[39] Itt most csupán a komponensek rövid ismertetésére szorítkozunk. A projekt gazdái az egyiptomi Biblioteca Alexandrina és az Új-Zélandi Nemzeti Könyvtár. Előbbi a technikai fejlesztésért, az utóbbi a kutatói, felhasználói igények becsatornázásáért felel elsősorban. A projekt három alapkomponensből áll. Az első egy link egy Link-indexer névre hallgató indexelő eszköz.[40] Ez kinyeri a szükséges metaadatokat a webarchívumokban tárolt WARC-fájlokból (WARC-fájl URI címe, WARC-fájl dátuma, az adott webhelyről kifelé mutató linkek listája), s egy önálló WAT-névre hallgató fájltípusban tárolja azokat. A Link-serv komponens a Link-indexer által kinyert adatokat szemantikus adattárban (data store) tárolja el, s gráf alapú adatsémát rendel hozzá. A gráf alapú adatbázist a Neo4j nevű NoSQL adatbáziskezelő rendszer menedzseli. A Link-Viz nevű harmadik komponens pedig egy megjelenítési felületet biztosít, ahol az adatbázisban tárolt adatok webböngészőn keresztül vizuálisan gráf-adatszerkezetben megjeleníthetők. Az egyes webhelyek egy-egy csomópontot alkotnak a hálózatban, a közöttük lévő kapcsolatok pedig térben és időben is tanulmányozhatóvá válnak.[41]

Remélhetőleg a jövőben egyre több hasonló projektről fogunk hallani, illetve a webarchiválás szolgáltatási környezetének megszilárdulása után, mi is szeretnénk Magyarországon kutatókkal együtt dolgozni big data alapú projekteken.

3. Hiteles webarchívum

A nemzeti könyvtárak hatókörén általában kívül eső komponens a webarchívumokból történő hitelesített adatok szolgáltatása. Az archivált hiteles jogi dokumentumok gyűjtése és felhasználása az üzleti és a közigazgatási szféra szintjén jelenik meg. Számos országban (például Nagy-Britanniában vagy Ausztráliában) törvény írja elő, hogy a cégek teljes online tevékenységét (ideértve a közönségkapcsolati csatornák forgalmát a közösségi médiafelületekkel együtt) archiválni kell, és a hiteles archivált anyagot jogi eljárásokban felhasználhatóvá kell tenni. Széles szakirodalma van a jogi hitelesség biztosításával kapcsolatos webarchiválási tevékenységnek[42] Egyes cégek erre építik fel üzleti modelljüket, hogy hiteles módon megőrzik a céges webes kommunikáció mindenféle lenyomatát, legyen szó weboldalról, vagy akár a közösségi médiáról. A brit MirrorWeb[43] cég például a tartalmi elemekben bekövetkező változásokat is rögzíti, naplózza. Így vissza lehet keresni, hogy egy jogi vita esetén adott konkrét időpontban milyen információkat bocsátott az adott cég ügyfeleinek rendelkezésére. A webhelyek mögött álló adatbázisok rekordjait napi szinten archiválják, s elérhetővé teszik audit, egyéb rendszeres ellenőrzési tevékenység, illetve ügyfélpanaszok kivizsgálásának céljából. Az államigazgatásban is egyre inkább előtérbe kerül néhány területen ez a kérdéskör, ahol a hiteles archivált anyagok szolgáltatása az államigazgatás átláthatósága, illetve a jogi viták eldöntése szempontjából jelenik meg. Itt most egy ausztrál példát említünk meg[44] .Természetesen ezt a hiteles másolatok begyűjtésére és kezelésére létrehozott teljes rendszerkörnyezetben lehet a leginkább megoldani. Ennek felépítése, a csatlakozó közigazgatási, pénzügyi szolgáltatások némelyikének áttekintése is megjelenik a szakirodalomban.[45] Új begyűjtési módszerek is előtérbe kerülnek ennek kapcsán.[46] Az üzleti élet és a közigazgatási szolgáltatások tisztességes és zavartalan működésének garantálásához ez a webarchiválást érintő terület várhatóan még jobban fel fog a jövőben értékelődni.

Epilógus

A webarchiválás mint átfogó interdiszciplináris szakmai kutatási terület egyre jobban intézményesül a nemzetközi tudományos életben. Ugyanez magyar viszonyok között még nem mondható el. Miután a webarchiválás mint szakmai feladat törvénymódosítás révén az Országos Széchényi Könyvtár alaptevékenységei között kap helyet, remélhetőleg, mint a tudományos vizsgálatok tárgya is szélesebb körben teret nyer majd a későbbiekben. A szerző amellett, hogy meg kívánta jelölni a személyes érdeklődése homlokterében álló kutatási területeket, egyben inspirációval is kíván szolgálni ahhoz, hogy minél többen, minél több nézőpontból válasszák a webarchiválást, a webarchívumot mint gyűjteményt a tudományos kutatásaik tárgyául.

Végezetül arról szeretnék szót ejteni, hogy az internet fejlődése számos nyitott kérdést rejt, melyek

megválaszolása alapvető hatással lehet arra, hogy a weben megjelenő tartalmakat miként lehetséges majd a jövőben archiválni. Megfigyelhető egyfajta egyre gyorsuló széttöredezettség, egyrészt a különféle platformok szintjén, másrészt a nagyhatalmi rivalizálás virtuális kivetüléseként. Ha a számítógépes világháló egységessége az eddigieknél még jobban háttérbe szorulna, az az archiválás szemszögéből új helyzetet teremthetne.

A webarchiválás mint üzleti tevékenységek tárgya eddig csak nagyon korlátozottan nyert teret. Ha e terület üzletileg esetleg felértékelődne a jövőben s nagy üzleti súlyú szereplők is megjelennének a szolgáltatásaikkal, az az eddig döntően a színtéren feltűnő közgyűjtemények, illetve egyéb nonprofit szereplők tevékenységi körének átértékelésével járhat majd.

Halvány elképzelésekkel rendelkezünk tehát arról, hogy mit hozhat a jövő. Egy azonban biztos. A Magyarországon 2017-ben elindult webarchiválási gyűjteményépítési tevékenység hamarosan szintet fog lépni, s a webarchívum mint gyűjtemény gyorsan bővülő anyaga remélhetőleg egyre több kutató érdeklődését kelti majd fel és sokszínű kutatási együttműködési lehetőségek kialakításának esélyét rejtheti magában.

Hivatkozások

[1] További információk a http://warcnet.eu oldalon érhetők el.

[2] Geeraert és Németh: Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary. https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_Hungary.pdf

[3] Brügger, Myrvoll, Schostag & Hunt: Exploring special web archive collections related to COVID-19: The case of Netarkivet. https://cc.au.dk/fileadmin/user_upload/WARCnet/Brügger_et_al_COVID-19_Netarkivet.pdf

[4] Geeraert and Bingham: Exploring special web archives collections related to COVID-19: The case of the UK Web Archive. https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_UKWA__1_.pdf

[5] Niels Brügger és mtsai., „Introduction: Internet histories”, Internet Histories 1, sz. 1–2 (2017. január 2.): 1–7, https://doi.org/10.1080/24701475.2017.1317128.

[6] Károly Kokas és László Drótos, „Webarchiválás és a történeti kutatások”, *Digitális Bölcsészet* 1, sz. 1 (2018. július 16.): 35–55, https://doi.org/10.31400/dh-hun.2018.1.129.

[7] László Drótos és Márton Németh, „Web museum, web library, web archive The responsibility of public collections to preserve digital culture”, in *The Power of Reading: Proceedings of the XXVI Bobcatsss Symposium, Riga, Latvia, January 2018*, szerk. Lelde Petrovska, Baiba Īvāne-Kronberga, és Zane Meldere (Riga: The University of Latvia Press., 2018), 124–26.

[8] Niels Brügger, „Introduction: The Web’s first 25 years”, *New Media & Society* 18, sz. 7 (2016. augusztus 8.): 1059–65, https://doi.org/10.1177/1461444816643787.

[9] Niels Brügger, „Digital Humanities in the 21st Century:Digital Material as a Driving Force”, *Digital Humanities Quarterly* 10, sz. 3 (2016), http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds.

[10] Niels Brügger, „Web historiography and Internet Studies: Challenges and perspectives”, *New Media & Society* 15, sz. 5 (2013. augusztus 21.): 752–64, https://doi.org/10.1177/1461444812462852.

[11] Susanne Belovari, „Historians and Web Archives.”, *Archivaria*, sz. 83 (2017): 59–79, http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds.

[12] Ada Lerner, Tadayoshi Kohno, és Franziska Roesner, „Rewriting History”, in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS ’17* (New York, New York, USA: ACM Press, 2017), 1741–55, https://doi.org/10.1145/3133956.3134042.

[13] Ahmed AlSum, „Reconstruction of the US First Website”, in *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL ’15* (New York, New York, USA: ACM Press, 2015), 285–86, https://doi.org/10.1145/2756406.2756954.

[14] Scott A Hale és mtsai., „Mapping the UK Webspace: Fifteen Years of British Universities on the Web”, in *Proceedings of the 2014 ACM Conference on Web Science*, WebSci ’14 (New York, NY, USA: ACM, 2014), 62–70, https://doi.org/10.1145/2615569.2615691.

[15] Daniel Gomes és mtsai., „Creating a billion-scale searchable web archive”, in *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion* (New York, New York, USA: ACM Press, 2013), 1059–66, https://doi.org/10.1145/2487788.2488118.

[16] Niels Brügger, „Website history and the website as an object of study”, *New Media & Society* 11, sz. 1–2 (2009. február): 115–32, https://doi.org/10.1177/1461444808099574.

[17] Justin F Brunelle és mtsai., „Not All Mementos Are Created Equal: Measuring the Impact of Missing Resources”, *International Journal on Digital Libraries* 16, sz. 3–4 (2015. szeptember): 283–301, http://dx.doi.org/10.1007/s00799-015-0150-6.

[18] Martin Klein, Harihar Shankar, és Herbert de Sompel, „Robust Links in Scholarly Communication”, in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18 (New York, NY, USA: ACM, 2018), 357–58, https://doi.org/10.1145/3197026.3203885;
Martin Klein, „The Memento Tracer Framework for Scalable High-Quality Web Archiving”, Presentation, 2019 International Internet Preservation Coalition General Assembly and Web Archiving Conference, June 5-7, 2019, Zagreb, Croatia, 2019, https://digital.library.unt.edu/ark:/67531/metadc1608967/.

[19] A legfontosabb ezek közül: Niels Brügger, „Web history and the website as an object of study”, New Media & Society 11, sz. 1-2 (2009): 115-132, https://doi.org/10.1177/1461444808099574

[20] Niels Brügger és Ditte Laursen, „A National Web Trend Index”, Presentation, 2019 International Internet Preservation Coalition General Assembly and Web Archiving Conference, June 5-7, 2019, Zagreb, Croatia, 2019. június 6., Denmark, https://digital.library.unt.edu/ark:/67531/metadc1608974/;
Janko Klasinc, „Web Archiving Overview: National and University Library - Slovenia”, Presentation, 2019 International Internet Preservation Coalition General Assembly and Web Archiving Conference, June 5-7, 2019, Zagreb, Croatia, 2019, Slovenia, https://digital.library.unt.edu/ark:/67531/metadc1609023/;
Karolina Holub, „Croatian Web Archive: practice and experiences in collecting Croatian web resources” (IIPC General Assembly, The Hague, The Netherlands, 2011. május 9.); Kees Teszelszky, „The harvest of the Dutch digital fields: the landscape of webarchiving in The Netherlands”, 2017, http://mekosztaly.oszk.hu/mia/doc/workshop/Kees_Teszelszky_2017_Presentatie_webarchivering_KB_BUDAPEST_404.ppt;
Kees Teszelszky, „Distant reading: The Frisian Web Domain”, 2019, http://mekosztaly.oszk.hu/mia/doc/DH_2019/2019.09.11_Teszelszky_Friese_web.pptx.

[21] Anat Ben-David, Adam Amram, és Ron Bekkerman, „The colors of the national Web: visual data analysis of the historical Yugoslav Web domain”, *International Journal on Digital Libraries* 19, sz. 1 (2018. március 18.): 95–106, https://doi.org/10.1007/s00799-016-0202-6.

[22] Brügger, „Website history and the website as an object of study”; Niels Brügger és Ralf Schroeder, szerk., *The Web as History: Using Web Archives to Understand the Past and the Present*, 1st kiad. (United States, North America: UCL Press, 2017), http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds.

[23] Brügger, „Website history and the website as an object of study”; Elisabetta Locatelli, „The role of Internet Wayback Machine in a multi-method research project”, in *“Researchers, pratictioners and their use of the archived web”, London, School of Advanced Study, University of London* (London, 2017).
„Asking questions with web archives – introductory notebooks for historians - IIPC”, 2020, http://netpreserve.org/projects/jupyter-notebooks-for-historians/.
A történészek lehetőségeiről lásd még: Ian Milligan: You shouldn't Need to be a Web Historian to Use Web Archives. https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf

[24] Ryan Deschamps, „Exploring Web Archival Data through Archives Unleashed Cloud Jupyter Notebooks”, Medium, 2019. március 12., https://news.archivesunleashed.org/exploring-web-archival-data-through-archives-unleashed-cloud-jupyter-notebooks-7605c6ca2b33;
Samantha Fritz és Ian Milligan, „Archive-It Blog – Analyze your Web Archives at Scale: The Archives Unleashed Cloud”, 2018, https://archive-it.org/blog/post/analyze-your-web-archives-at-scale-the-archives-unleashed-cloud/.

[25] „Jupyter notebooks for web archives”, 2020, https://slides.com/wragge/iipc-jupyter.

[26] „Jupyter notebooks for web archives”, 2020, https://slides.com/wragge/iipc-jupyter.

[27] „Asking questions with web archives – introductory notebooks for historians - IIPC”.

[28] „Jupyter notebooks for web archives”.

[29] „Jupyter notebooks for web archives".

[30] „Final report. Asking questions with web archives - Introductory notebooks for historians", 2020, http://netpreserve.org/projects/jupyter-notebooks-for-historians/.

[31] Emily Maemura, Christoph Becker, és Ian Milligan, „Understanding computational web archives research methods using research objects", in *2016 IEEE International Conference on Big Data (Big Data)* (IEEE, 2016), 3250–59, https://doi.org/10.1109/BigData.2016.7840982;
Helge Holzmann, Wolfram Sperber, és Mila Runnwerth, „Archiving Software Surrogates on the Web for Future Reference.", *Research & Advanced Technology for Digital Libraries: 20th International Conference on Theory & Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings*, 2016. január, 215, http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds;
Helge Holzmann, Wolfgang Nejdl, és Avishek Anand, „On the Applicability of Delicious for Temporal Search on Web Archives", in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16* (New York, New York, USA: ACM Press, 2016), 929–32, https://doi.org/10.1145/2911451.2914724.

[32] Martin Lnenicka, Jan Hovad, és Jitka Komarkova, „A Proposal of a Big Web Data Application and Archive for the Distributed Data Processing with Apache Hadoop", in *Computational Collective Intelligence. Lecture Notes in Computer Science, vol 9330*, szerk. Manuel Núñez és mtsai. (Cham: Springer International Publishing, 2015), 285–94, https://doi.org/10.1007/978-3-319-24306-1_28.

[33] Lnenicka, Hovad, és Komarkova.

[34] Lnenicka, Hovad, és Komarkova.

[35] Brügger és Schroeder, *The Web as History: Using Web Archives to Understand the Past and the Present*; Márton Németh, „A webarchiválásról történeti megközelítésben", *Könyv, könyvtár, könyvtáros* 27, sz. 2 (2018):48–52, http://ki2.oszk.hu/3k/2018/06/a-webarchivalasrol-torteneti-megkozelitesben/.

[36] Jane Winters, „Big UK Domain Data for the Arts and Humanities", Presentation, 2015 International Internet Preservation Coalition General Assembly, April 27 - May 1, 2015. Silicon Valley, California., 2015. április 27., https://digital.library.unt.edu/ark:/67531/metadc1476406/;
Jane Winters; Josh Cowls, „Research Using Big UK Domain Data", Presentation, 2015 International Internet Preservation Coalition General Assembly, April 27 - May 1, 2015. Silicon Valley, California., 2015. április 27., https://digital.library.unt.edu/ark:/67531/metadc1476399/.

[37] WEB Archive UK és JICS, „Shine Project Historical Research Prototype", 2015, https://www.webarchive.org.uk/shine.

[38] WEB Archive UK és JICS.

[39] „LinkGate: Core Functionality and Future Use Cases - IIPC", 2020, http://netpreserve.org/projects/LinkGate/.

[40] „IIPC RSS LinkGate Webinar", 2020, https://docs.google.com/presentation/d/1mYSciOvbU9Hm3jsMSJgioSVr3ZGuVkVyr10HnHXJwXA/edit#slide=id.g8cec5e7f8a_0_67.

[41] „IIPC RSS LinkGate Webinar".

[42] Néhány példa csupán: G. Patrick Flanagan, „Digital Preservation and Authentic Legal Information", SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2010), https://doi.org/10.2139/ssrn.2463288;
Jennie Grimshaw, „UK Official Publications: Managing the Transition to Electronic Deposit at the British Library", *Legal Information Management* 16, sz. 1 (2016. március): 3–9, http://dx.doi.org/10.1017/S1472669616000037;
Jason Webber, „Using Secondary Datasets for Researchers under a Legal Deposit Framework", Presentation, 2019 International Internet Preservation Coalition General Assembly and Web Archiving Conference, June 5-7, 2019, Zagreb, Croatia, 2019. június 6., United Kingdom, https://digital.library.unt.edu/ark:/67531/metadc1608986/.

[43] „Website Archiving and Monitoring Solutions | MirrorWeb", 2020, https://www.mirrorweb.com.

[44] Flanagan, „Digital Preservation and Authentic Legal Information". Social Science Research Network. 2010. https://papers.ssrn.com/abstract=2463288

[45] S Thornton, „Value and impact: Third Northumbria international conference on performance measurements in libraries and information services", *Managing Information* 6, sz. 9 (1999): 89; Mihai Togan és Ionut Florea, „A Reference Model for a Trusted Service Guaranteeing Web-Content", in *ISSE 2015*, szerk. Helmut Reimer, Norbert Pohlmann, és Wolf-

gang Schneider (Wiesbaden: Springer Fachmedien Wiesbaden, 2015), 216–24, https://doi.org/10.1007/978-3-658-10934-9_18.

[46] Sawood Alam és mtsai., „Supporting Web Archiving via Web Packaging", IAB 2019, 3. https://www.iab.org/wp-content/IAB-uploads/2019/06/sawood-alam-2.pdf

Beérkezett: 2020. XI. 30-án.

Németh Márton
Országos Széchényi Könyvtár
Információ és Tartalomszolgáltatási Webarchiválási Osztály.
E-mail: nemeth.marton@oszk.hu
URL: http://webarchivum.oszk.hu/